ECE 408

Fall 2019

Final Project

Milestone 2

Team Name: happy_parallel

Team member 1:Hanyue Gu NetID: hanyueg2 UIN:677049370

Team member 2: Zhongxiu Xie NetID: zx27 UIN:655299530

Team member 3: Yichi Zhang NetID: yichi3 UIN:661294886

Report Deliverables:

1. List of all kernels that collectively consume more than 90% of the program time.
2. List of all CUDA API calls that collectively consume more than 90% of the program time.
3. Explanation of the difference between kernels and API calls.
4. Output of rai running MXNet on the CPU.
5. List program run time.
6. Output of rai running MXNet on the GPU.
7. List program run time.
8. List whole program execution time.
9. List Op Times.

1. kernels that collectively consume more than 90% of the program time

Kernel Name	Time (ms)	Time (%)
[CUDA memcpy HtoD]	35.929	33.17
volta_scudnn_128x64_relu_interior_nn_v1	17.992	16.61
volta_gcgemm_64x32_nt	17.388	16.05
fft2d_c2r_32x32	9.6060	8.87
volta_sgemm_128x128_tn	7.8609	7.26
op_generic_tensor_kernel	7.3342	6.77
fft2d_r2c_32x32	7.1221	6.58
Total count:	103.2322	95.31

2. CUDA API calls that collectively consume more than 90% of the program time

Kernel Name	Time (s)	Time (%)
cudaStreamCreateWithFlags	3.13539	43.19
cudaMemGetInfo	2.26800	31.24
cudaFree	1.58861	21.89
Total count:	6.99200	96.32

3. Difference between kernels and API calls
   Kernels are user defined code that are run on the GPU devices and the execution units are threads that within kernel blocks and streaming processors on GPU. On the other hand, API calls are predefined and users need to call these API functions in host code in order to make connections between device and host. For example, we use “cudaMalloc” to allocate memory space in device for the variables that we want to handle.

4. Output of rai running MXNet on the CPU

```
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8154}
17.52user 4.53system 0:09.02elapsed 244%CPU (0avgtext+0avgdata 6047576maxresiden
t)k
0inputs+2824outputs (0major+1601669m
inor)pagefaults 0swaps
```

5. Running time: 17.52s on user, 4.53 on system and 9.02s of elapsed time.
6. Output of rai running MXNet on the GPU

```
* Running /usr/bin/time python m1.2.py
Loading fashion-mnist data... done
Loading model... done
New Inference
EvalMetric: {'accuracy': 0.8154}
5.19user 3.03system 0:06.16elapsed 133%CPU (0avgtext+0avgdata 2987800maxresident)k
0inputs+4536outputs (0major+733425minor)pagefaults 0swaps
```

7. Running time: 5.19s on user, 3.03s on system and 6.16s of elapsed time.

8. Whole program execution time (CPU)

```
* Running /usr/bin/time python m2.1.py
Loading fashion-mnist data... done
Loading model... done
New Inference
Op Time: 11.465068
Op Time: 63.306210
Correctness: 0.7653 Model: ece408
90.49user 11.11system 1:18.92elapsed 128%CPU (0avgtext+0avgdata 6042312maxresident)k
0inputs+0outputs (0major+2307358minor
)pagefaults 0swaps
```

   90.49s on user, 11.11s on system and 1 min 18.92 s of elapsed time.

9. Op Time1: 11.465068s
   Op Time2: 63.306210s
   Correctness: 0.7653